

Incremental Learning of Event Definitions with Inductive Logic Programming

Nikos Katzouris · George Paliouras ·
Alexander Artikis

Received: date / Accepted: date

Abstract Event Recognition systems rely on properly engineered knowledge bases of event definitions to infer occurrences of events in time. The manual development of such knowledge is a tedious and error-prone task, thus event-based applications may benefit from automated knowledge construction techniques, such as Inductive Logic Programming (ILP), which combines machine learning with the declarative and formal semantics of First-Order Logic. However, learning temporal logical formalisms, which are typically utilized by logic-based Event Recognition systems is a challenging task, which most ILP systems cannot fully undertake. In addition, event-based data are usually massive and collected at different times and under various circumstances. Ideally, systems that learn from temporal data should be able to operate in an incremental mode, that is, revise prior constructed knowledge in the face of new evidence. Most ILP systems are batch learners, in the sense that in order to account for new evidence they have no alternative but to forget past knowledge and learn from scratch. Given the increased inherent complexity of ILP and the volumes of real-life temporal data, this results to algorithms that scale poorly. In this work we present an incremental method for learning and revising event-based knowledge, in the form of Event Calculus programs. The proposed algorithm relies on abductive-inductive learning and comprises an scalable clause refinement methodology, based on a compressive summarization of clause coverage in a stream of examples. We present an empirical evaluation of our approach on real and synthetic data from a video surveillance application.

1 Introduction

The growing amounts of temporal data collected during the execution of various tasks within organizations are hard to utilize without the assistance of automated processes. Event Recognition (Etzion and Niblett, 2010; Luckham, 2001; Luckham and Schulte, 2008) refers to the automatic detection of event occurrences within

Institute of Informatics & Telecommunications, National Center of Scientific Research “Demokritos”

a system. From a sequence of *low-level events* (for example sensor data) an event recognition system recognizes *high-level events* of interest, that is, events that satisfy some pattern. Event recognition systems with a logic-based representation of event definitions, such as the Event Calculus (Kowalski and Sergot, 1986), are attracting significant attention in the event processing community for a number of reasons, including the expressiveness and understandability of the formalized knowledge, their declarative, formal semantics (Paschke, 2005; Artikis et al, 2010a) and their ability to handle rich background knowledge. Using logic programs in particular, has an extra advantage, due to the close connection between logic programming and machine learning in the field of Inductive Logic Programming (ILP) (Muggleton and Raedt, 1994; Lavrac and Dzeroski, 1993). However, such applications impose challenges that make most ILP systems inappropriate.

Several logical formalisms which incorporate time and change employ non-monotonic operators as a means for representing commonsense phenomena (Mueller, 2006). Normal logic programs (Lloyd, 1987) with Negation as Failure (NaF) (Clark, 1977) in particular are a prominent non-monotonic formalism. Most ILP learners cannot handle NaF at all, or lack a robust NaF semantics (Sakama, 2000; Ray, 2009). Another problem that often arises in temporal reasoning, is the need to infer implicit or missing knowledge, for instance the indirect effects of events, or possible causes of observed events. In ILP the ability to reason with missing, or indirectly observable knowledge is called *non-Observational Predicate Learning (non-OPL)* (Muggleton, 1995). This is a task that most ILP systems have difficulty to handle, especially when combined with NaF in the background knowledge (Ray, 2006). One way to address this problem is through the combination of ILP with Abductive Logic Programming (ALP) (Denecker and Kakas, 2002; Kakas and Mancarella, 1990; Kakas et al, 1993). Abduction in logic programming is usually given a non-monotonic semantics (Eshghi and Kowalski, 1989) and in addition, it is by nature an appropriate framework for reasoning with incomplete knowledge. Although it has a long history in the literature (Ade and Denecker, 1995), only recently has this combination brought about systems such as XHAIL (Ray, 2009), TAL (Corapi et al, 2010) and ASPAL (Corapi et al, 2011b; Athakravi et al, 2013) that may be used for the induction of event-based knowledge.

The above three systems which, to the best of our knowledge, are the only ILP learners that address the aforementioned learnability issues are *batch learners*, in the sense that all training data must be in place prior to the initiation of the learning process. This is not always suitable for event-oriented learning tasks, where data are often collected at different times and under various circumstances. In order to account for new incoming training examples a batch learner has no alternative but to re-learn a hypothesis from scratch. This comes at the cost of increased overhead and poor scalability when “learning in the large” (Dietterich et al, 2008) from a growing set of data. This is particularly true in the case of temporal data, which usually come at large volumes. Consider for instance data which span a large period of time, or sensor data transmitted at a very high frequency.

An alternative approach is learning incrementally that is, processing training instances when they become available, and altering previously inferred knowledge to fit new observations, instead of discarding it and starting from scratch. This process, also known as *Theory Revision* (Wrobel, 1996), exploits previous computations to speed-up the learning, since revising a hypothesis is generally considered

more efficient than learning it from scratch (Biba et al, 2008; Esposito et al, 2000; Cattafi et al, 2010). Numerous theory revision systems have been proposed in the literature – see (Esposito et al, 2000) for a review — however their applicability in non-monotonic domains is limited (Corapi et al, 2008). This issue is addressed by more recent approaches to *theory revision as non-monotonic ILP* (Corapi et al, 2008; Maggi et al, 2011; Corapi et al, 2011a), where a non-monotonic learner is used to extract a set of prescriptions, which can in turn be interpreted into a set of syntactic transformations on the theory at hand. Scalability is a known drawback of theory revision, and the development of scalable algorithms has been identified as an important research direction (Muggleton et al, 2012b). As historical data grow over time, it becomes progressively harder to revise knowledge, so that it accounts both for new evidence and past experience. One direction towards scaling theory revision systems would be the development of techniques for reducing the need for reconsulting the whole history of accumulated experience, while updating existing knowledge.

In this work we use XHAIL as a basis for the development of a scalable, incremental learner for the induction of event definitions in the form of Event Calculus theories. XHAIL has been used for the induction of action theories (Sloman and Lupu, 2010; Alrajeh et al, 2010, 2011, 2012, 2009). Moreover, in (Corapi et al, 2008) it has been used for theory revision in an incremental setting, revising hypotheses with respect to a recent, user-defined subset of the perceived experience. In contrast, the learner we present here performs revisions that account for all examples seen so far. We describe a bottom clause-based, compressive “memory” structure, incorporated in the learning process, which reduces the need for reconsulting past experience in response to a revision. In particular, we describe a method which, given a stream of examples, a theory which accounts for them and a new training instance, requires one pass over the examples in order to revise the initial theory, so that it accounts for both past and new evidence, under certain locality assumptions. We evaluate empirically our approach on real and synthetic data from a video surveillance application.

The rest of this paper is structured as follows. Section 2 focuses on the Event Calculus dialect that we employ. In Section 3 we describe the domain of activity recognition where our examples are drawn from. In Section 4 we review abductive-inductive learning and the XHAIL system, while in Section 5.2 we present our proposed method, prove its correctness and describe its abductive-inductive implementation. In Section 7 we present experimental evaluation, and finally in Sections 8 and 9 we discuss related work and summarize our main conclusions.

2 Event Calculus in Logic Programming

In this section we review some basic notions of Logic Programming, as well as the Event Calculus. Let us assume a first-order language, where terms, term substitutions, atoms, literals, clauses, integrity constraints and logic programs are defined in the usual way (Lloyd, 1987), and `not` denotes NaF. Following Prolog’s convention, predicates and ground terms start with a lower case letter and variable terms start with a capital letter. A logic program is *Horn* if it contains no negated literals and *normal* otherwise. An atom α θ -subsumes an atom β , denoted $\alpha \preceq \beta$, if there exists a substitution θ such that $\alpha\theta = \beta$. A clause C θ -subsumes a clause D ,

Predicate	Meaning
$\text{happens}(E, T)$	Event E occurs at time T
$\text{initiatedAt}(F, T)$	At time T a period of time for which fluent F holds is initiated
$\text{terminatedAt}(F, T)$	At time T a period of time for which fluent F holds is terminated
$\text{holdsAt}(F, T)$	Fluent F holds at time T

Table 1: The basic predicates in the language of the SDEC

denoted $C \preceq D$, if there exists a substitution θ such that $\text{head}(C)\theta = \text{head}(D)$ and $\text{body}(C)\theta \subseteq \text{body}(D)$. Finally a program P_1 θ -subsumes a program P_2 if for each clause $C \in P_1$ there exists a clause $D \in P_2$ such that $C \preceq D$.

Given a logic program P and an interpretation I , that is, a subset of the set of all possible groundings of P , we say that I satisfies a literal α (resp. not α) iff $\alpha \in I$ (resp. $\alpha \notin I$). I satisfies a set of ground atoms iff it satisfies each one of them and it satisfies a ground clause iff it satisfies the head, or does not satisfy at least one body literal. I is a model of P iff it satisfies every ground instance of every clause in P and it is a minimal model iff no strict subset of I is a model of P . Finally I is a *stable model* of P iff it is a minimal model of the Horn program that results from the ground instances of P after the removal of all clauses with a negated literal not satisfied by I , and all negative literals from the remaining clauses. P entails a set of ground literals E (den. $P \models E$) iff at least one stable model of P satisfies E .

The Event Calculus (Kowalski and Sergot, 1986) is a temporal logic for reasoning about events and their effects. It has been used for event recognition in (Artikis et al, 2010a). The ontology of the Event Calculus comprises *time points*, i.e integer of real numbers; *fluents*, i.e properties which have certain values in time; and *events*, i.e occurrences in time that may affect fluents and alter their value. In this work we assume that fluents are boolean-valued and time points are positive integers. The core, domain-independent axioms of the formalism incorporate the common sense *law of inertia*, according to which fluents persist over time, unless they are affected by an event. We call the Event Calculus dialect used in this work Simplified Discrete Event Calculus (SDEC). As its name implies, it is a simplified version of the Discrete Event Calculus, a dialect which is equivalent to the classical Event Calculus when time ranges over integer domains (Mueller, 2008).

The building blocks of the SDEC are presented in Table 1. The core, domain-independent axioms of the SDEC are:

$$\begin{array}{ll} \text{holdsAt}(F, T+1) \leftarrow \text{initiatedAt}(F, T). & (1) \end{array} \quad \begin{array}{ll} \text{holdsAt}(F, T+1) \leftarrow \text{holdsAt}(F, T), \\ \text{not terminatedAt}(F, T). & (2) \end{array}$$

Axiom (1) states that a fluent F holds at time T if it has been initiated at the previous time point, while Axiom (2) states that F continues to hold unless it is terminated. $\text{initiatedAt}/2$ and $\text{terminatedAt}/2$ are defined in an application-specific manner. Examples are presented in the section that follows.

Narrative	Annotation
.....
<code>happens(<i>inactive</i>(<i>id</i>₁), 999)</code>	
<code>happens(<i>active</i>(<i>id</i>₂), 999)</code>	
<code>holdsAt(<i>coords</i>(<i>id</i>₁, 201, 432), 999)</code>	<code>not holdsAt(<i>moving</i>(<i>id</i>₁, <i>id</i>₂), 999)</code>
<code>holdsAt(<i>coords</i>(<i>id</i>₂, 230, 460), 999)</code>	<code>not holdsAt(<i>moving</i>(<i>id</i>₂, <i>id</i>₁), 999)</code>
<code>holdsAt(<i>direction</i>(<i>id</i>₁, 280), 999)</code>	
<code>holdsAt(<i>direction</i>(<i>id</i>₂, 270), 999)</code>	
<code>happens(<i>walking</i>(<i>id</i>₁), 1000)</code>	
<code>happens(<i>walking</i>(<i>id</i>₂), 1000)</code>	
<code>holdsAt(<i>coords</i>(<i>id</i>₁, 201, 454), 1000)</code>	<code>not holdsAt(<i>moving</i>(<i>id</i>₁, <i>id</i>₂), 1000)</code>
<code>holdsAt(<i>coords</i>(<i>id</i>₂, 230, 440), 1000)</code>	<code>not holdsAt(<i>moving</i>(<i>id</i>₂, <i>id</i>₁), 1000)</code>
<code>holdsAt(<i>direction</i>(<i>id</i>₁, 270), 1000)</code>	
<code>holdsAt(<i>direction</i>(<i>id</i>₂, 270), 1000)</code>	
<code>happens(<i>walking</i>(<i>id</i>₁), 1001)</code>	
<code>happens(<i>walking</i>(<i>id</i>₂), 1001)</code>	
<code>holdsAt(<i>coords</i>(<i>id</i>₁, 201, 454), 1001)</code>	<code>holdsAt(<i>moving</i>(<i>id</i>₁, <i>id</i>₂), 1001)</code>
<code>holdsAt(<i>coords</i>(<i>id</i>₂, 227, 440), 1001)</code>	<code>holdsAt(<i>moving</i>(<i>id</i>₂, <i>id</i>₁), 1001)</code>
<code>holdsAt(<i>direction</i>(<i>id</i>₁, 275), 1001)</code>	
<code>holdsAt(<i>direction</i>(<i>id</i>₂, 278), 1001)</code>	
.....

Table 2: An annotated stream of LLE's

3 Running example: Activity recognition

Throughout this paper we use the task of video surveillance, as defined in the CAVIAR¹ project, as a running example. The CAVIAR dataset consists of videos of a public space, where actors walk around, meet each other, browse information displays, fight and so on. These videos have been manually annotated by the CAVIAR team, to provide the ground truth for two types of activity. The first type corresponds to low-level events, that is, knowledge about a person's activities at a certain time point (for instance *walking*, *running*, *standing still* and so on). The second type corresponds to high-level events, activities that involve more than one person, for instance two people *moving together*, *fighting*, *meeting* and so on. The aim is to recognize high-level events by means of combinations of low-level ones and some additional domain knowledge, such as a person's position and direction at a certain time point.

Low-level events are represented in SDEC by streams of ground `happens/2` atoms (see Table 2), while high-level events and other domain knowledge are represented by ground `holdsAt/2` atoms. Streams of low-level events together with domain-specific knowledge will henceforth constitute the *narrative*, in ILP terminology, while knowledge about high-level events is the *annotation*. Table 2 presents an annotated stream of low-level events. We can see for instance that the person *id*₁ is *inactive* at time 999, her (*x*, *y*) coordinates are (254, 756) and her direction is 270°. The annotation for the same time point informs us that *id*₁ and *id*₂ are not moving together. Given that both target high-level events (for instance *moving*) and some low-level ones (for instance *coords*, see Table 2) are considered fluents by being arguments to a `holdsAt` predicate, we discriminate between *inertial* and

¹ <http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/>

statically defined fluents. The former should be inferred by the Event Calculus axioms, while the latter are provided with the input.

Given such a domain description in the language of the SDEC, the aim of machine learning addressed in this work is to automatically derive the *Domain-Specific Axioms*, that is, the axioms that specify the way in which the occurrence of low-level events affect the truth values of the fluents that represent high-level events, by initiating or terminating them. Thus, we wish to learn *initiatedAt/2* and *terminatedAt/2* definitions. The learnability issues mentioned in the Introduction are evident from this example: The learning problem we just described is non-OPL, since instances of target predicates are not provided with the supervision. Furthermore, the abductive derivation of the missing instances is combined with NaF, via the axioms of SDEC, and this makes this task impossible for most ILP systems.

4 Abductive-Inductive learning

Given logic programs B (background knowledge), E (examples or observations) and M (hypothesis language), an ILP algorithm constructs a program H (inductive hypothesis) in the language defined by M , such that $B \cup H \models E$, in which case we also say that $B \cup H$ *explains* or *covers* E . In this paper examples are sets of ground literals, while B consists of the axioms of SDEC, unless otherwise stated, and H is a logic program in the language of SDEC. The hypothesis language (or syntactic bias) provides general guidelines concerning the structure of literals that may appear in the head or the body of a hypothesis clause, the types of their arguments, and the way variables are linked to each other. A commonly used syntactic bias in ILP, also employed in this work, is *mode declarations* (Muggleton, 1995).

Mode declarations define the *schema* for the allowed literals in the clause, and can be of the form *modeh(s)* or *modeb(s)* for the head or body literals respectively. The schema s is a ground “template” literal with *input*, *output* and *ground* placemarkers, which are special terms of the form $+\text{type}$, $-\text{type}$ and $\#\text{type}$ respectively, where *type* corresponds to a term’s type. The distinction between input and output terms is that any input term in a body literal must be an input term in the head, or an output term in some preceding body literal. A set M of mode declarations defines a language $\mathcal{L}(M)$ in which clauses are constructed by defining head and body literals via the *modeh* and *modeb* directives respectively. Literals are constructed from the schemas in M as follows: by replacing an output placemaker by a new variable; by replacing an input placemaker by a variable that appears in the head literal, or in a previous body literal; and by replacing a ground placemaker by a ground term.

Abduction derives the conditions under which a set of observations can be made true. Abductive Logic Programming (ALP) is formalized as a triplet $ALP(B, A, E)$ where B is some logic theory, E is a set of ground literals (observations) and A is a set of predicates called *abducibles*. A set of ground atoms Δ is called an explanation of the ALP task if $B \cup \Delta \models E$ and the predicate of each atom in Δ appears in A . A number of implementations that allow abductive reasoning (Ray and Kakas, 2006; Mancarella et al, 2009) exist. Moreover, reasoners based on Answer Set Programming (ASP) (Marek and Truszczyński, 1999; Niemela, 1999),

called answer set solvers, are highly efficient tools that naturally support abduction via built-in constructs. A large number of tasks involved in learning with ILP can be practically carried out by an ASP solver. As a result, the implementation of the presented work is based almost exclusively on ASP.

4.1 XHAIL

XHAIL is an ILP system that interleaves abductive and inductive reasoning in a three-phase process. Given examples E , background knowledge B and mode declarations M , XHAIL's first two phases return a ground program K in the language of M , called *Kernel Set of E and B* , which entails E with respect to B . The first phase generates the heads of the Kernel Set clauses by abductively deriving from B instances of head mode declaration atoms that satisfy E . The second phase generates the Kernel Set, by saturating each previously abduced atom with instances of body declaration atoms that deductively follow from $B \cup \Delta$. Next, the Kernel Set is variabilized, that is, each term that corresponds to a variable, as indicated by the mode declarations, is replaced by an actual variable (identical terms are replaced by the same variable). XHAIL's third phase searches the space of theories that θ -subsume the variabilized Kernel Set, in order to find a hypothesis H such that $B \cup H \models E$. The search bias is *minimality*, that is, preference towards the smallest (in terms of overall number of literals) theory that covers all positive and no negative examples.

To give an example, XHAIL will abductively derive two atoms $\Delta = \{\text{initiatedAt}(\text{moving}(id_1, id_2), 1000), \text{initiatedAt}(\text{moving}(id_2, id_1), 1000)\}$ in its first phase, from the knowledge presented in Table 2, by using appropriate mode declarations (see Table 3). Next, it will add to each abduced atom literals derivable from $B \cup \Delta$ resulting in a Kernel Set, as presented in Table 3.

In its third phase, XHAIL generalizes the (variabilized) kernel set to a minimal hypothesis, by dropping as many body literals and clauses as possible. Search is also implemented abductively. The variabilized Kernel Set is first transformed by Algorithm 1 to a program U_K , in order to allow the abductive derivation of a minimal set of $use/2$ atoms from the abductive task $T = ALP(B \cup U_K, \{use/2\}, E)$. The intuition is that in order for clause C_i in the Kernel Set to cover an example from E , each of its body literals β_i^j must be true. The transformation applied to C_i offers two choices for that, either prove β_i^j (in which case $use(i, j)$ must also be abduced), or assume $\text{not } use(i, j)$. Thus the presence of a $use(i, j)$ atom in an explanation of T indicates that the j -th literal of the i -th Kernel Set clause contributes towards example coverage. $use(i, 0)$ corresponds to the head literal of the i -th clause. On the other hand, the minimality bias exploits the second option for proving β_i^j (i.e., assuming $\text{not } use(i, j)$) as much as possible, thus including in an abductive explanation only the $use/2$ instances which correspond to literals that are necessary to explain E . Any literal that lacks a corresponding $use/2$ atom in the abductive explanation may be discarded. One of the possible minimal hypotheses thus generated is H , shown in Table 3. It consists of a most-general clause that explains the examples in Table 2.

Note that in Table 3, most mode declarations are omitted for simplicity and only a fragment of the body literals in each Kernel Set clause (which depend on the employed body declarations) is presented. For instance, mode declarations include

Mode declarations	Abduced atoms (XHAIL's phase 1)
$modeh(\text{initiatedAt}(\text{moving}(+person, +person), +time))$ $modeb(\text{happens}(\text{walking}(+person), +time))$ $modeb(\text{holdsAt}(\text{close}(+person, +person, \#distance), +time))$	$\text{initiatedAt}(\text{moving}(id_1, id_2), 1000)$ $\text{initiatedAt}(\text{moving}(id_2, id_1), 1000)$
A Kernel Set (XHAIL's phase 2):	
$\text{initiatedAt}(\text{moving}(id_1, id_2), 1000) \leftarrow$ $\text{happens}(\text{walking}(id_1), 1000),$ $\text{happens}(\text{walking}(id_2), 1000),$ $\text{holdsAt}(\text{close}(id_1, id_2, 33), 1000).$	$\text{initiatedAt}(\text{moving}(id_2, id_1), 1000) \leftarrow$ $\text{happens}(\text{walking}(id_2), 1000),$ $\text{happens}(\text{walking}(id_1), 1000),$ $\text{holdsAt}(\text{close}(id_1, id_2, 33), 1000).$
Preparation for searching the variabilized Kernel Set (Algorithm 1):	
$\text{initiatedAt}(\text{moving}(X, Y), T) \leftarrow$ $\text{use}(1, 0), \text{try}(1, 1, [X, T]),$ $\text{try}(1, 2, [Y, T]), \text{try}(1, 3, [X, Y, T]).$ $\text{try}(1, 1, [X, T]) \leftarrow \text{not use}(1, 1).$ $\text{try}(1, 1, [X, T]) \leftarrow \text{use}(1, 1),$ $\text{happens}(\text{walking}(X), T).$ $\text{try}(1, 2, [Y, T]) \leftarrow \text{not use}(1, 2).$ $\text{try}(1, 2, [Y, T]) \leftarrow \text{use}(1, 2),$ $\text{happens}(\text{walking}(Y), T).$ $\text{try}(1, 3, [X, Y, T]) \leftarrow \text{not use}(1, 3).$ $\text{try}(1, 3, [X, Y, T]) \leftarrow \text{use}(1, 3),$ $\text{holdsAt}(\text{close}(X, Y, 33), T).$	$\text{initiatedAt}(\text{moving}(X, Y), T) \leftarrow$ $\text{use}(2, 0), \text{try}(2, 1, [X, T]),$ $\text{try}(2, 2, [Y, T]), \text{try}(2, 3, [X, Y, T]).$ $\text{try}(2, 1, [X, T]) \leftarrow \text{not use}(2, 1).$ $\text{try}(2, 1, [X, T]) \leftarrow \text{use}(2, 1),$ $\text{happens}(\text{walking}(X), T).$ $\text{try}(2, 2, [Y, T]) \leftarrow \text{not use}(2, 2).$ $\text{try}(2, 2, [Y, T]) \leftarrow \text{use}(2, 2),$ $\text{happens}(\text{walking}(Y), T).$ $\text{try}(2, 3, [X, Y, T]) \leftarrow \text{not use}(2, 3).$ $\text{try}(2, 3, [X, Y, T]) \leftarrow \text{use}(2, 3),$ $\text{holdsAt}(\text{close}(X, Y, 33), T).$
Abductive solution (XHAIL's phase 3)	Hypothesis
$\Delta = \{\text{use}(1, 0), \text{use}(1, 3)\}$	$H = \text{initiatedAt}(\text{moving}(X, Y), T) \leftarrow$ $\text{holdsAt}(\text{close}(X, Y, 33), T).$

Table 3: Computation of an inductive hypothesis H from the knowledge in Table 2 and SDEC. $\text{close}(X, Y, D)$ is a (statically defined) fluent which represents the Euclidean distance D between persons X and Y . For simplicity, we present only a number of the actually employed mode declarations and only parts of the generated Kernel Set clauses.

Algorithm 1 Transformation of a variabilized Kernel Set for abductive search (Ray, 2009)

- 1: **let** K be a *variabilized* Kernel Set, i.e, the program that results by replacing each term in a Kernel Set clause that corresponds to a variable (as indicated by the mode declarations), by a variable.
- 2: **let** $U_K = \emptyset$
- 3: **for each** clause $C_i = \alpha_i \leftarrow \beta_i^1, \dots, \beta_i^n \in K$ **do**
- 4: **let** $U_K = U_K \cup \{\alpha_i \leftarrow \text{use}(i, 0), \text{try}(i, 1, \text{vars}_i^1), \dots, \text{try}(i, n, \text{vars}_i^n)\}$, where vars_i^j
- 5: is a term that contains the variables that appear in literal β_i^j
- 6: **for each** literal $\beta_i^j \in \text{body}(C_i)$ **do**
- 7: **let** $U_K = U_K \cup \{\text{try}(i, j, \text{vars}_i^j) \leftarrow \text{use}(i, j), \beta_i^j\} \cup \{\text{try}(i, j, \text{vars}_i^j) \leftarrow \text{not use}(i, j)\}$
- 8: **return** U_K .

several other head atom specifications for different CAVIAR high-level events in addition to *moving* (*meeting*, *fighting*, *leaving an object* etc). They also include body declarations for other CAVIAR low-level events in addition to *walking* (*active*, *inactive*, *abrupt*, *running* etc.) and also body declarations for the generation of negated body literals via Closed World Assumption.

The Kernel Set is a non-monotonic, multi-clause version of the *Bottom Set*, a concept widely used by prominent Inverse Entailment ILP systems like PROGOL (Muggleton, 1995) and ALEPH². Due to the property of θ -subsuming a kernel set of E and B , XHAIL’s proof procedure is called *Kernel Set Subsumption* and an XHAIL-generated hypothesis H for the ILP task $ILP(B, M, E)$ is said to be *derivable by Kernel Set Subsumption from E and B* . Contrary to set covering algorithms used by PROGOL and ALEPH, where positive examples are covered in turn until no more are left uncovered, XHAIL explains all the examples in one go, by generalizing all clauses in a Kernel Set. This has a number of advantages (Ray, 2009, 2006): (a) It ensures soundness, i.e if a hypothesis is returned, then it covers all the provided examples, something that does not hold in general for standard set-covering approaches in the non-monotonic setting. (b) It leads to optimal hypothesis (in terms of the size of the hypothesis), in contrast to set covering approaches, where the selection of locally optimal hypotheses may result to a globally sub-optimal one. (c) It successfully handles certain cases of inherent incompleteness, i.e inability to compute all possible hypotheses, that characterize Inverse Entailment-based algorithms. On the other hand, this approach is naturally not-scalable, due to the increased combinatorial complexity of the search space. The trade-off for a tractable search space is to relax requirements for (a), (b) and (c) above, and use XHAIL in a set covering loop, as indicated in (Ray, 2006). In such a setting, each example e is covered in turn by generalizing a Kernel Set of e and B .

Either as a “theory-level” induction system, or as a set-covering algorithm, XHAIL is a batch learner, as is the majority of ILP systems. The training set must be in place prior to the initiation of the learning and in case new training instances are presented, previously induced knowledge cannot be utilized. Instead, it must be discarded and a new hypothesis must be generated from the new, augmented training set.

5 Incremental Learning with ILED

In this section we present our approach towards an incremental learner for the induction of Event Calculus programs, based on XHAIL. The goal is two-fold: First, utilize any previously induced knowledge as a revisable background theory, which may be used in order to cover new incoming training instances. If no such theory is present, the learner should be able to induce one from scratch. Second, revisions should be done efficiently, provided that they account for the whole set of past experience.

Since the target theories in this work comprise event definitions in the form of Event Calculus rules, we discriminate between a “fixed” background theory (the axioms of SDEC) and a revisable background theory, that is, sets of Domain-Specific Axioms that have been previously induced. Henceforth, we use the term “example” to encompass anything known true at a specific time point, plus negated annotation literals at that time point, obtained by closed world assumption. An example’s time point will also serve as reference. For instance, three different examples (e_{999} , e_{1000} and e_{1001}) are presented in Table 2. According to the annotation,

² <http://web.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph/>

an example is either positive or negative w.r.t a particular high-level event. For instance, e_{1000} in Table 2 is a negative example for the *moving* high-level event, while e_{1001} is a positive example. In this setting the task of incremental learning addressed in this work can be described as follows:

Definition 1 (Incremental Learning) Let \mathcal{E} be a (possibly empty) database of examples, called historical memory, B the axioms of SDEC, M a set of mode declarations, H a (possibly empty) hypothesis such that $B \cup H \models \mathcal{E}$ and an incoming training instance $w \notin \mathcal{E}$. The goal is to revise H to a theory H_{rev} such that $B \cup H_{rev} \models \mathcal{E} \cup w$.

Maintaining a complete historical memory of the examples to which the learner has been exposed, allows to implement revisions which account for all past experience. On the other hand, this experience may grow over time to an extent that is impossible to maintain in the working memory. We thus follow an *external memory* approach (Biba et al, 2008) and implement \mathcal{E} as an external database to which the learner has access. This design decision implies that past experience is never processed as a whole, and thus each set of examples fetched from the database should be explainable independently of the other examples. Note for instance, that the knowledge presented in Table 2 is explainable (as shown in Section 4), independently of any other past or future examples. On the contrary, this would not hold for a set of examples explained by a clause of the form

$$\begin{aligned} \text{initiatedAt}(\text{moving}(X, Y), T) \leftarrow \\ \text{happens}(\text{walking}(Y), T1), \\ T1 < T. \end{aligned}$$

since, in order to explain what happens at T , knowledge about what happens at some unspecified $T1$ prior to T is required. Such independence assumptions are called *locality assumptions* in the *Learning from Interpretations* (Blockeel et al, 1999) ILP setting. Learning from Interpretations is considered appropriate for most event-based applications (Dubba et al, 2011) and is the setting that we adopt in this work. When Learning from Interpretations, each example is a set of facts that may be considered as a separate database which can be queried independently (Blockeel et al, 1999). Definition 2 makes the locality assumption more specific in the context of this work.

Definition 2 (Locality assumption) Let \mathcal{E} and M be as in Definition 1. The locality assumption holds for \mathcal{E} iff (a) there exists a non-empty hypothesis H in the language of M , derivable by Kernel Set Subsumption from \mathcal{E} and SDEC; (b) for each mode declaration $m \in M$ the schema of M does not contain any output variable placemarkers of type **time**.

Condition (a) serves as a “no noise” assumption for the example set \mathcal{E} . Throughout this work by “noise” we mean contradictions or missing knowledge in the training data. Condition (b) ensures that each example $e \in \mathcal{E}$ may be explained independently using the axioms of SDEC and the knowledge in e only. In particular, under condition (b) the examples may be considered as database entries with **time** being the primary key. In addition to example independence in a set of data, condition (b) also results in a constrained search space when constructing hypotheses with Kernel Set Subsumption. This is because the number of literals that appear in the body of a Kernel Set clause is constrained by their time stamp,

which should be referenced in the head literal. This results in a smaller space of theories that subsume the Kernel Set, even if NaF is allowed in the body of Kernel Set clauses.

To be precise, knowledge about a specific time point alone does not suffice to explain the example at this time point, as the axioms of SDEC make the truth value of a fluent at time T dependable on what happens at $T - 1$. Thus, we assume that the input is given in the form of *sliding windows*, i.e sets of at least two temporally successive examples. For instance, the data in Table 2 may be considered as part of two windows, or as part of a single window. The size of a window w (i.e the number of examples it contains) is measured by the *granularity* G of the window. For instance, if the data in Table 2 are considered as part of two windows then each window has $G = 2$. For a single window which contains all the data it holds that $G = 3$.

5.1 Revising hypotheses

Assume a hypothesis H and an example window w . We say that H covers w , denoted by $B \cup H \models w$, if $B \supset H$ covers all positive and none of the negative examples in w . If $B \cup H \not\models w$, then H is either *incomplete*, *inconsistent*, or both. A hypothesis is incomplete when it does not cover some positive examples and inconsistent when it covers some negatives. If a hypothesis is both complete and consistent then we say that it is *correct*. A clause is consistent when it covers no negatives and it is correct if it is consistent and it does not disprove any positives.

A Theory Revision algorithm employs *generalization* and *specialization* (*refinement*) operators to act upon a theory and alter its *answer set* (Wrobel, 1996; Esposito et al, 2000), that is, the examples it accounts for. Generalization operators aim to increase example coverage and may add new clauses or remove literals from existing clauses. Specialization operators aim to restrict example coverage and may remove clauses or add antecedents to existing clauses. A Theory Revision algorithm typically begins by identifying “failing points” (Wrobel, 1996) in the theory at hand, and responds by generalization whenever the theory does not cover positive examples (incompleteness) and specialization when it covers negatives (inconsistency). However, this strategy is not generally applicable in non-monotonic domains. For instance, clause addition, while typically a generalization, may restrict the coverage of the whole theory, as in the case of adding a *terminatedAt* clause in a hypothesis. Similarly, removal of clauses from a hypothesis, or their specialization, may increase the coverage of the theory, as shown in Example 1.

Example 1 Consider the following knowledge related to the *fighting* high-level event from CAVIAR:

Narrative :	Annotation :
<code>happens(abrupt(id₁), 1).</code>	<code>not holdsAt(fighting(id₁, id₂), 1).</code>
<code>happens(abrupt(id₂), 1).</code>	<code>not holdsAt(fighting(id₂, id₁), 1).</code>
<code>holdsAt(close(id₁, id₂, 23), 1).</code>	<code>holdsAt(fighting(id₁, id₂), 2).</code>
<code>happens(walking(id₁), 2).</code>	<code>holdsAt(fighting(id₂, id₁), 2).</code>
<code>happens(abrupt(id₂), 2).</code>	<code>holdsAt(fighting(id₁, id₂), 3).</code>
<code>holdsAt(close(id₁, id₂, 23), 2).</code>	<code>holdsAt(fighting(id₂, id₁), 3).</code>

Hypothesis :

$$\begin{aligned}
 C_1 = & \text{initiatedAt}(\text{fighting}(X, Y), T) \leftarrow \\
 & \text{happens}(\text{abrupt}(X), T), \\
 & \text{not happens}(\text{inactive}(Y), T), \\
 & \text{holdsAt}(\text{close}(X, Y, 23), T). \\
 C_2 = & \text{terminatedAt}(\text{fighting}(X, Y), T) \leftarrow \\
 & \text{happens}(\text{walking}(X), T).
 \end{aligned}$$

Clause C_1 states that *fighting* between two persons id_1 and id_2 is initiated if one of them exhibits an *abrupt* behavior and the other is not *inactive*. Clause C_2 states that *fighting* is terminated if someone walks. Clause C_1 is consistent w.r.t the presented knowledge and it does not disprove any positives, thus it is correct. On the other hand, clause C_2 disproves the positive example $\text{holdsAt}(\text{fighting}(id_1, id_2), 3)$, thus it is incorrect. As a result, the hypothesis $H = C_1 \cup C_2$ is not satisfied by the knowledge presented above. Since H fails to cover a positive example it is incomplete. Specializing C_2 to clause

$$\begin{aligned}
 C'_2 = & \text{terminatedAt}(\text{fighting}(X, Y), T) \leftarrow \\
 & \text{happens}(\text{walking}(X), T), \\
 & \text{not holdsAt}(\text{close}(X, Y, 23), T).
 \end{aligned}$$

lifts the incompleteness, thus a specialization increases the coverage of the hypothesis. Clause C'_2 now dictates that *fighting* between two persons is terminated when one of the involved persons walks away.

□

5.2 ILED

In this section we describe our approach to Incremental Learning of Event Definitions (ILED). ILED's input and goal are described in Definition 1. Once a new window $w \notin \mathcal{E}$ is presented, ILED checks whether w satisfies $B \cup H$. If not, it begins the process of revising H in order to account for $\mathcal{E} \cup w$. In this process, revision operators that retract knowledge, such as the deletion of clauses or antecedents are excluded, due to the exponential cost of backtracking in the historical memory (Badea, 2001). The allowable revision operators supported by ILED are:

- Addition of *initiatedAt* clauses or specialization of *terminatedAt* clauses to treat incompleteness.
- Addition of *terminatedAt* clauses or specialization of *initiatedAt* clauses to treat inconsistency.

Clause addition is based on XHAIL's Kernel Set generalization and ILED reduces to XHAIL when $\mathcal{E} = H = \emptyset$. In that case, an initial hypothesis is obtained via generalizing a Kernel Set generated from B and the incoming training window w , which is subsequently added to \mathcal{E} . We next describe our approach to clause refinement in ILED.

As mentioned earlier, the historical memory \mathcal{E} is implemented as an external database which grows over time. It is never entirely loaded into the working memory, as we assume that from a point on, it simply does not “fit”. Thus reconsulting \mathcal{E} in order to restore completeness and consistency after a revision is done sequentially, using one window at a time. Under this setting, we measure the complexity

Algorithm 2 Support set construction and update

```

let  $w \notin \mathcal{E}$  be a new training instance and  $C$  a clause
if  $C$  is a new clause then
  if  $C$  is generated by KSS from a (variabilized) Kernel Set  $K$  of  $w$  then
     $C.\text{supp} \leftarrow \{D \in K \mid C \preceq D\}$ 
  if  $C$  results from the refinement of clause  $C.\text{parent}$  then
     $C.\text{supp} \leftarrow \{D \in C.\text{parent}.\text{supp} \mid C \preceq D\}$ 
if  $C$  is an existing clause then
  let  $e_C^w$  the true positives that  $C$  covers in  $w$ , if  $C$  is an initiatedAt clause, else
   $e_C^w$  be the true negatives that  $C$  covers, if it is a terminatedAt clause.
  if  $C.\text{supp}$  does not cover  $e_C^w$  then
     $C.\text{supp} \leftarrow C.\text{supp} \cup K'$ , where  $K'$  is the subset of a (variabilized) Kernel Set
    of  $w$ , which covers  $e_C^w$ 

```

of ILED by the number of “passes” over the entire database \mathcal{E} that are required in order to revise H .

Incremental learning as described in Definition 1 is expensive in principle. For instance, assume that H is generalized by the addition of a new *initiatedAt* clause in order to increase example coverage w.r.t the incoming training window w . Then \mathcal{E} must be checked for consistency since the new clause may cover negatives. In this case, the inconsistent clause should be refined, which in turn may affect its initial coverage in w . New refinements must then be generated and added to H , and these refinements should be again checked for consistency in \mathcal{E} . This process of successive generalizations and specializations continues until a revised hypothesis that accounts for all the examples in $\mathcal{E} \cup w$ is found. As illustrated above, in general this requires several passes over the entire historical memory.

To address this issue ILED employs a compressive memory structure, the *support set*, which aims to facilitate clause refinement by providing an effective clause refinement bias, using bottom clauses as a pool for potential antecedents. The support set is merely a set of variabilized Kernel Set clauses. It is a clause-level structure, that is, it is associated to each theory clause C generated by ILED. The support set of a clause C , denoted by $C.\text{supp}$, is generated once C itself is generated, and that may happen in two ways, as shown in Algorithm 2: (a) C is an entirely new clause, generated by Kernel Set Subsumption from a Kernel Set K of some window w . Then $C.\text{supp}$ initially contains the (variabilized) Kernel clauses from K , which are θ -subsumed by C . (b) C is generated from an existing clause $C.\text{parent}$ by the addition of a set of antecedents. In this case $C.\text{supp}$ initially contains the clauses of $C.\text{parent}.\text{supp}$ which continue to be θ -subsumed by C . The support set of an existing theory clause C is “updated” as new windows are added in the historical memory after being processed by ILED, as shown in Algorithm 2, so that $C.\text{supp}$ always summarizes the examples that C covers throughout \mathcal{E} , while abstracting away the redundant parts of the search space. Due to this property, the support set may be utilized for scaling up the clause refinement process, since in order to refine clause C while ensuring that the refinement preserves the original coverage of clause C in the historical memory, it suffices to replace C by a program R_C that θ -subsumes $C.\text{supp}$. In what follows we call such clause refinements *supported refinements*.

Thanks to the support set, ILED is able to implement a revision strategy which allows to save a lot of redundant inference effort in the historical memory

— see Figure 1. In this figure, example windows are represented by circles, while hypotheses are represented by squares. Arrows that connect a hypothesis with a window (or a set of windows), indicate that the hypothesis is correct (complete & consistent), or possibly incorrect (incomplete or inconsistent) w.r.t that window. Figure 1 presents the historical memory as a set of example windows w_1, \dots, w_n and an initial hypothesis H , which is complete & consistent w.r.t the historical memory, as in Definition 1. A new incoming window instance w_{n+1} is provided. Figure 1 presents two cases that may occur:

1. A new clause must be added to H in response to w_{n+1} , resulting in hypothesis H_1 (Figure 1(a)).
2. A new literal must be added to an existing clause in H in response to w_{n+1} , resulting in hypothesis H_1 (Figure 1(b)).

In the first case, the new clause in H_1 may be an `initiatedAt` clause which account for the positives that H fails to cover in w_{n+1} , or it may be a `terminatedAt` clause in order to reject negatives that H covers in w_{n+1} . In case of an `initiatedAt` clause, H_1 may cover negatives in the historical memory and in case of a `terminatedAt` clause H_1 may disprove positives. This is indicated by an “incorrectness” arrow that connects H_1 to the whole historical memory in Figure 1(a). In response, each window in the historical memory is checked for completeness & consistency, and H_1 is further revised if necessary.

For instance, assume that the first window where H_1 is incorrect is w_{n-1} (then $H_2 = H_1$, since no revisions were necessary w.r.t window w_n). Assume that the incorrectness is due to a newly added `initiatedAt` clause $C \in H_1$, which covers some negatives in w_{n-1} . By means of the support set, clause C will be replaced by a program R_C which θ -subsumes $C.\text{supp}$ and thus preserves the initial coverage of clause C in window w_{n+1} . The situation is similar in case C is a `terminatedAt` clause which prevents some positives in w_{n-1} to be covered. The hypothesis H_3 in Figure 1(a) which results from the supported refinement of clause C is complete and consistent w.r.t the new window w_{n+1} , which does not need to be re-checked. It is also complete & consistent w.r.t window w_n . This is because the revision which yields H_3 consists of the replacement of clause $C \in H_1$ by a set R_C of refinements of C . Then if C is an `initiatedAt` clause, it does not cover negatives in w_n , since we assumed that w_{n-1} is the first window where H_1 is incorrect. Therefore none of the refinements in the program R_C which replaces C cover negatives in w_n . Similarly it follows that C does not disprove any positives in w_n in case it is a `terminatedAt` clause.

As the “correctness” arrows in Figure 1(a) indicate, hypothesis H_3 is complete and consistent w.r.t all the windows which have been checked so far, namely w_{n+1}, w_n, w_{n-1} . It follows inductively that as the windows in the historical memory are sequentially checked for completeness and consistency, the implemented revisions correctly account for all the windows which have already been checked. That is, the hypothesis H_{n-j} is complete & consistent w.r.t window w_{n-j+1} , for each $j \leq n+1$, thus none of these windows need to be rechecked. Thus, for $j = n+1$ (i.e when all n windows in the historical memory, plus the new window w_{n+1} have been checked) the resulting hypothesis is complete & consistent w.r.t $\mathcal{E} \cup w_{n+1}$ (as required), and moreover each example window in the historical memory has been checked exactly once.

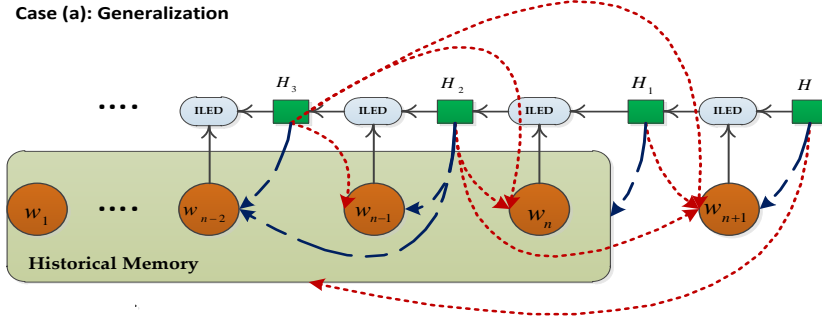
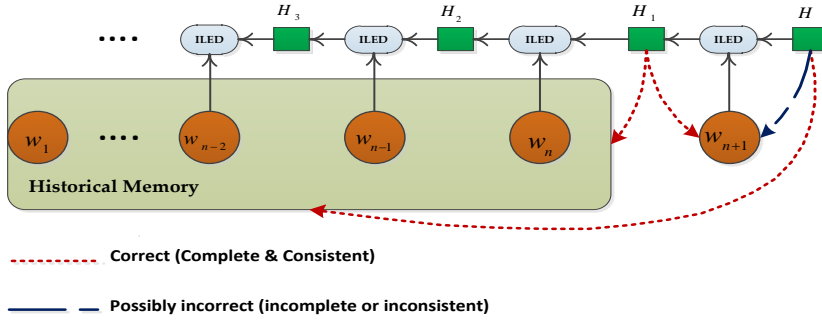
Case (a): Generalization**Case (b): Specialization**

Fig. 1: ILED revision strategy.

In the second case mentioned above, hypothesis H_1 results from the initial hypothesis H by the supported refinement of an existing clause (Figure 1(b)). Since each such supported refinement preserves the coverage of the initial parent clause in the historical memory, H_1 correctly accounts for all experience and no further actions are required, as indicated by the “correctness” arrows in Figure 1(b).

In general, a hypothesis may be both incomplete and inconsistent w.r.t to the incoming window. ILED tries refinements of existing clauses by adding antecedents drawn from support clauses. At the same time it generalizes new Kernel clauses, constructed by the examples in the incoming window, in order to induce new clauses which account for uncovered examples. The implemented revisions are selected based on minimality, that is, preference towards the smallest revision which is satisfied by the incoming window. However, these revisions are not guided exclusively by the knowledge in the current window, but also have the property of maximizing the preserved prior coverage by means of the support set. The implementation of the general strategy relies on Abductive Logic Programming. A revision is constructed by an abductive search in the space of clauses that subsume a new Kernel Set and at the same time, in the space of antecedents defined by the support clauses.

ILED’s single revision operator is presented in Algorithm 3. The algorithm accepts three types of input program:

Algorithm 3 `revise`($w, B, M, \text{Correct}_w, \text{Incorrect}_w, \text{Kernel}_w$)

Input: A window w , background knowledge B , mode declarations M and programs $\text{Correct}_w, \text{Incorrect}_w, \text{Kernel}_w$ in the language of M .

Output: A tuple $\langle \text{Correct}'_w, R_w, K_w \rangle$, where $\text{Correct}'_w$ is a consistent program which does not disprove positives, R_w is a program that subsumes the support of each clause in the input Incorrect_w program and K_w is a set of clauses that θ -subsume the input Kernel_w program.

```

1: let  $T_{inc}$  be the program that results by applying consistency_transformation to each
   clause in  $\text{Incorrect}_w$ 
2: let  $T_{kernel}$  be the program that results by applying kernel_transformation to each clause
   in  $\text{Kernel}_w$ 
3: let  $\Phi$  be the abductive task  $\Phi = ALP(B, \text{Cons}_w \cup T_{inc} \cup T_{kernel}, \{use/2, use/3\}, w)$ 
4: if  $\Phi$  has a solution then
5:   let  $\Delta$  be a minimal solution of  $\Phi$ 
6:   let  $K_w$  be the program obtained by removing from the  $i$ -th clause of  $\text{Kernel}_w$  the
       $j$ -th body literal for each  $use(i, j) \notin \Delta$ , and removing the  $i$ -th Kernel clause for
      each  $use(i, j) \notin \Delta$ 
7:   let  $\text{Correct}'_w$  be the program which contains the  $i$ -th clause of  $\text{Incorrect}_w$  if  $use(i, j, k) \notin \Delta$ 
8:   let  $Ref$  be the program obtained by adding to the  $i$ -th clause of  $\text{Incorrect}_w \setminus \text{Correct}'_w$  the
       $k$ -th body literal of its  $j$ -th support clause, for each atom  $use(i, j, k) \in \Delta$ 
9:   for each clause  $C_i \in \text{Incorrect}_w \setminus \text{Correct}'_w$  do
10:    let  $T_{C_i} = \text{consistency\_transformation}(C_i)$ 
11:    let  $\Delta'$  be the set of all minimal solutions of the abductive task
       $\Phi' = ALP(B, T_{C_i} \cup (Ref \setminus C_i) \cup K_w \cup \text{Correct}'_w, \{use/3\}, w)$ 
12:    let  $ref(C_i)$  be the set of all minimal refinements of  $C_i$  which may be
      constructed with literals from the solutions in  $\Delta'$ , as in line 8
13:    let  $R_{C_i}$  be the smallest subset of  $ref(C_i)$  which  $\theta$ -subsumes  $C_i.supp$ 
14:    let  $R_w = \bigcup_{C_i \in \text{Incorrect}_w \setminus \text{Correct}'_w} R_{C_i}$ 
15: else
16:   Return No Solution
17: Return  $\langle \text{Cons}'_w, R_w, K_w \rangle$ 
18:
19: Subroutine consistency_transformation( $C_i$ )
20: Let  $T_{C_i}$  be the program obtained as follows:
21: add to  $T_{C_i}$  the clause  $head(C_i) \leftarrow body(C_i) \wedge \text{not } expt(i, head(C_i)')$ 
22: where  $head(C_i)'$  carries the variables that appear in  $head(C_i)$ 
23: for each clause  $\Gamma_i^j \in C_i.supp$  do
24:   add to  $T_{C_i}$  one clause  $expt(i, head(C_i)') \leftarrow use(i, j, k), \text{not } \beta_{i,j}^k, \forall \beta_{i,j}^k \in body(\Gamma_i^j)$ 
25: Return  $T_{C_i}$ 
26:
27: Subroutine generate_refinements $C_i, \Gamma_i^j$ 
28:
29: Subroutine kernel_transformation( $C_i$ ) : Algorithm 1 applied to  $\text{Kernel}_w$ 

```

- (a) A (potentially empty) set of clauses which we know they are correct, that is, each such clause covers no negatives and it does not disprove any positives. These clauses are meant to be retained during the revision step.
- (b) A potentially incorrect set of clauses. During the revision step, each of these clauses may be replaced by a correct program, consisting of supported refinements of the initial incorrect clause.

- (c) A set of Kernel Set clauses. These clauses are generated by the examples of the current window and, during the revision step, they are used for the induction of new clauses if necessary, in order to cover new examples.

The Kernel set clauses and the potentially incorrect ones are syntactically transformed respectively into programs that allow for an abductive search in the space of clauses that subsume the Kernel Set, and the space of clauses that subsume the support set of each potentially incorrect clause. Solutions to this abductive search may be interpreted as actual instructions for the construction of clauses that subsume a Kernel Set clause, so that new examples may be covered, or for the addition of antecedents to incorrect clauses. For the (variabilized) Kernel Set clauses, this syntactic transformation is presented in Algorithm 1 and discussed in Section 4.1. For potentially incorrect clauses, the transformation is obtained by means of the `consistency.transformation` (see Algorithm 3).

The `consistency.transformation` is based on an exception learning technique, widely used in non-monotonic ILP. To refine a clause $p \leftarrow q$ one writes it as $p \leftarrow q \wedge \text{not } \text{expt}$, where *expt* stands for an “exception” predicate, and learns a definition for *expt*. The revised clause is generated by adding one literal from each clause in the definition of the exception predicate. The consistency transformation utilized in Algorithm 3 uses the same principle, but skips the learning part. Instead, it explicitly provides a set of possible definitions for the exception predicate, namely one for each literal in the support of the clause under refinement. The goal is then to select a minimal set of support literals which collectively cause the clause under refinement to fail. This is achieved by means of the *use/3* abducibles, which once abducted, may be used as prescriptions for the addition of antecedents. More details are provided in Example 2 below.

Given an incorrect clause C_i , Algorithm 3 searches the space of antecedents defined by each clause in $C_i.\text{supp}$, in order to find a set of refinements which θ -subsume the whole support set, since this ensures that the program which will replace C_i preserves the original coverage of the clause. This is implemented by a set cover search in the subsets of all supported refinements of clause C_i , which returns the smallest subset that θ -subsumes the support set.

Altering the input to the `revise` function, obtains different specifications that suit different tasks. When a new window w is presented to the learner and fails to satisfy the hypothesis at hand, the `revise` function is called with specification `revise($w, B, \emptyset, H, \text{Kernel}_w$)`, where Kernel_w is a Kernel Set generated from the examples in w . In the above specification, we consider all clauses in the current hypothesis as potentially incorrect, thus the incorrect part in ILED’s input amounts to the whole H . A different specification is used when ILED deals with a window from the historical memory. Assume for instance that a new clause C was generated in response to window w_{n+1} in Figure 1(a) and the algorithm tests window $w_n \in \mathcal{E}$ for completeness and consistency. Then the specification `revise($w_n, B, H \setminus C, C, \emptyset$)` is employed, which identifies C as the only potentially incorrect clause and omits a Kernel Set, since potential incompleteness and inconsistency may be lifted by the refinement of the new clause C .

The search bias towards minimal revisions is implemented by appropriate prescriptions to the abductive reasoner to minimize the number of atoms that will be abducted. This bias ensures that the implemented revisions are absolutely necessary. In particular, no *use/2* atoms indicating Kernel literals that cover a number

Hypothesis	
Clauses C_1 and C_2 from Example 1	
Support set ($C_2.supp$):	
$C_2^1 = \text{terminatedAt}(\text{fighting}(X, Y), T) \leftarrow$ $\quad \text{happens}(\text{walking}(X), T),$ $\quad \text{happens}(\text{standing}(Y), T),$ $\quad \text{not holdsAt}(\text{close}(X, Y, 23), T),$ $\quad \dots\dots$	$C_2^2 = \text{terminatedAt}(\text{fighting}(X, Y), T) \leftarrow$ $\quad \text{happens}(\text{walking}(X), T),$ $\quad \text{happens}(\text{running}(Y), T),$ $\quad \text{not holdsAt}(\text{close}(X, Y, 23), T),$ $\quad \dots\dots$
Example window w :	
Narrative & annotation from Example 1	
Program T_{C_2} (application of consistency_transformation on clause C_2):	
$\text{terminatedAt}(\text{fighting}(X, Y), T) \leftarrow$ $\quad \text{happens}(\text{walking}(X, T), T),$ $\quad \text{not expt}(1, [X, Y, T]).$ $\text{expt}(1, [X, Y, T]) \leftarrow$ $\quad \text{use}(1, 1, 1), \text{not happens}(\text{standing}(Y), T).$ $\text{expt}(1, [X, Y, T]) \leftarrow$ $\quad \text{use}(1, 1, 2), \text{holdsAt}(\text{close}(X, Y, 23), T).$ $\dots\dots$	$\text{terminatedAt}(\text{fighting}(X, Y), T) \leftarrow$ $\quad \text{happens}(\text{walking}(X, T), T),$ $\quad \text{not expt}(2, [X, Y, T]).$ $\text{expt}(2, [X, Y, T]) \leftarrow$ $\quad \text{use}(1, 2, 1), \text{not happens}(\text{running}(Y), T).$ $\text{expt}(2, [X, Y, T]) \leftarrow$ $\quad \text{use}(1, 2, 2), \text{holdsAt}(\text{close}(X, Y, 23), T).$ $\dots\dots$
(some) abductive solutions	Resulting refinements
$\Delta_1 = \{\text{use}(1, 1, 1), \text{use}(1, 2, 1)\}$	$C_{ref}^1 = \text{terminatedAt}(\text{fighting}(X, Y), T) \leftarrow$ $\quad \text{happens}(\text{walking}(X, Y), T),$ $\quad \text{happens}(\text{standing}(Y), T).$ $C_{ref}^2 = \text{terminatedAt}(\text{fighting}(X, Y), T) \leftarrow$ $\quad \text{happens}(\text{walking}(X), T),$ $\quad \text{happens}(\text{running}(Y), T).$
$\Delta_2 = \{\text{use}(1, 1, 2), \text{use}(1, 2, 2)\}$	$C_{ref}^3 = \text{terminatedAt}(\text{moving}(X, Y), T) \leftarrow$ $\quad \text{happens}(\text{walking}(X), T),$ $\quad \text{not holdsAt}(\text{close}(X, Y, 23), T).$ $C_{ref}^4 = \text{identical to } C_{ref}^3$
$\dots\dots$	$\dots\dots$
Optimal refinement:	
C_{ref}^3 (since $C_{ref}^3 \preceq C_2^1$ and $C_{ref}^3 \preceq C_2^2$)	

Table 4: Clause refinement in ILED

of examples will be abduced, in case the input hypothesis covers all examples in an incoming window w . Similarly, no $use/3$ atoms will be abduced, indicating support literals that reject examples, if the corresponding clauses in the input hypothesis are consistent.

To sum up, ILED works by progressively augmenting an initially empty hypothesis with *partial* clauses, that is, minimal clauses that may be incorrect w.r.t the historical memory or w.r.t future examples. These clauses are refined whenever necessary in order to cover positive examples or reject negative examples. By means of the support set, these refinements are ensured to lift local incompleteness or inconsistency, while preserving the global prior coverage of the clause, thus reducing the need for reconsulting past experience. The proposed setting is tailored towards incremental learning, where new examples may arrive during the learning process. Moreover, it may also be utilized in a batch setting, where all examples are available from the start, towards more scalable learning.

Example 2 Table 4 presents ILED's revision process. Recall that in Example 1 clause C_2 is incorrect since it disproves a positive example. In contrast, clause

C_1 is correct. The goal is for ILED to preserve the correct clause, while refining the incorrect one. We assume that the knowledge (narrative & annotation) from Example 1 is a new incoming window and, as a result, all existing clauses in the hypothesis at hand will be treated as potentially incorrect. Thus they will both be analysed according to the **consistency_transformation** in order to facilitate the search in the space of antecedents defined by their support sets. To save space, in Table 4 we present the results of applying the transformation only to clause C_2 . We also omit the new Kernel Set constructed from the new window, from which new clauses may be potentially induced in order to account for uncovered examples. The process of constructing a Kernel Set is presented in Table 3.

We assume that clause C_2 has two support clauses, C_2^1 and C_2^2 , as shown in Table 4. For brevity, we only show a fragment of each support set, relevant to the revision task presented in this example. The result of applying **consistency_transformation** on clause C_2 for each one of its support clauses is the program T_{C_2} . This program (in addition to a corresponding program T_{C_1} generated by clause C_1 and the new Kernel Set, analysed by Algorithm 1) is input to an abductive reasoner, together with the background theory and the examples in w as goals. The intuition is that in order for C_2 to fail so that the uncovered positive example is covered, the complement of each exception literal must fail, thus the exception literal itself must succeed. This may be achieved by disproving one of the support literals and abducing the corresponding *use/3* atoms. Each abduced *use*(i, j, k) is to be interpreted as a prescription of the form “use the k -th literal of the j -th support clause of the i -th theory clause”. The complement of each literal indicated by an abduced atom in a minimal solution of the abduction is added to C_2 , thus generating the corresponding refinement. Both solutions presented in Table 4 result in an acceptable set of refinements, that is, a set of correct clauses which subsume the support set. However, the best refinement is C_{ref}^3 , which results from solution Δ_2 , since it θ -subsumes both support clauses. This refinement covers the single positive which previously the hypothesis $C_1 \cup C_2$ failed to cover. All other positives are correctly accounted for by clause C_1 , which moreover is consistent. This means that no minimal abductive solution will contain *use/3* prescriptions for the refinement of clause C_1 , since it is not necessary. Clause C_1 is thus returned as is. Similarly, as all examples are covered after the refinement, there is no need to introduce new clauses, thus a minimal abductive solution will contain no *use/2* prescriptions from which new clauses that subsume the Kernel Set may be generated. The new hypothesis consists of clause C_1 , returned unchanged, and the refined clause $C_2' = C_{ref}^3$.

□

6 Discussion

In this section we discuss some properties and shortcomings of ILED and describe some directions for future extensions of the presented work. The trade-off for efficiency is that not all of ILED’s revisions are fully evaluated on the historical memory. In particular, a new clause generated by Kernel Set Subsumption in response to an incoming window w is selected among a set of possible choices, based on the examples that the clause covers at w . The negative examples it covers in the historical memory (in case it is an *initiatedAt* clause), or the positive examples it

disproves (in case it is a `terminatedAt` clause) are not taken into account. This may result in a large number of refinements and an unnecessarily lengthy hypothesis, as compared to one that may have been obtained by selecting a different initial clause. On the other hand, selecting an optimal set of new clauses which correctly cover all the examples in the incoming training instance, while minimizing coverage of unwanted examples throughout \mathcal{E} requires extensive inference in \mathcal{E} . Thus optimality in ILED has been sacrificed for efficiency.

In ILED, a large part of the theorem proving effort that is involved in clause refinement reduces to computing subsumption between clauses, which is a hard task. Moreover, just as the historical memory grows over time, so do (in the general case) the support sets of the clauses in the running hypothesis, increasing the cost of computing subsumption. However, as in principle the largest part of a search space is redundant and the support set focuses only on its interesting parts, one would not expect that the support set will grow to a size that makes subsumption computation less efficient than inference over the entire \mathcal{E} . Moreover, the locality assumption from Definition 2 helps in restricting the size of support clauses and makes the computation of subsumption relations tractable. In addition, a number of optimization techniques have been developed over the years and several generic subsumption engines have been proposed (Maloberti and Sebag, 2004; Kuzelka and Zelezny, 2008; Santos and Muggleton, 2010), some of which are able to efficiently compute subsumption relations between clauses comprising thousands of literals and hundreds of distinct variables.

ILED is oriented towards soundness, meaning that if a hypothesis is returned, then it covers all examples currently in the historical memory. This imposes some restrictions on its application domain. In particular, we assume that the supervision is complete and correct (i.e it contains no contradictions or missing knowledge) and the domain is stationary, in the sense that knowledge already induced remains true with respect to future instances, and retracting clauses or literals from the hypothesis at hand is never *necessary* in order to account for new incoming example windows. ILED terminates in case its computations result in a dead-end, returning `No Solution` (see Algorithm 3), which indicates that the only hypothesis that can account for the whole set of examples seen so far, is a plain enumeration of the examples. This results in treating cases widely studied in incremental learning, such as *concept drift* (Esposito et al, 2004), as noise. It is possible to relax the requirement for soundness and aim at an implementation that best-fits the training instances. Handling noise and concept drift are promising extensions of ILED.

7 Experimental evaluation

In this section we present experimental results on real data from the benchmark CAVIAR video surveillance dataset³, as well as large volumes of synthetic data.

³ <http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/>

7.1 Experimental setting

Recall that in CAVIAR the goal is to learn definitions of high-level events, such as *fighting*, *moving*, *meeting*, from streams of low-level events like *walking*, *standing*, *active*, *abrupt*, as well as spatio-temporal knowledge. Details on the CAVIAR dataset are presented in (Artikis et al, 2010b). Consider for instance the following definition of the *fighting* high-level event:

$$\begin{aligned} \text{initiatedAt}(\text{fighting}(X, Y), T) \leftarrow \\ \text{happens}(\text{active}(X), T), \\ \text{not happens}(\text{inactive}(Y), T), \\ \text{holdsAt}(\text{close}(X, Y, 23), T). \end{aligned} \quad (3)$$

$$\begin{aligned} \text{initiatedAt}(\text{fighting}(X, Y), T) \leftarrow \\ \text{happens}(\text{abrupt}(X), T), \\ \text{not happens}(\text{inactive}(Y), T), \\ \text{holdsAt}(\text{close}(X, Y, 23), T). \end{aligned} \quad (4)$$

$$\begin{aligned} \text{terminatedAt}(\text{fighting}(X, Y), T) \leftarrow \\ \text{happens}(\text{walking}(X), T), \\ \text{not holdsAt}(\text{close}(X, Y, 23), T). \end{aligned} \quad (5)$$

$$\begin{aligned} \text{terminatedAt}(\text{fighting}(X, Y), T) \leftarrow \\ \text{happens}(\text{running}(X), T), \\ \text{not holdsAt}(\text{close}(X, Y, 23), T). \end{aligned} \quad (6)$$

Rule (3) dictates that a period of time for which two persons X and Y are assumed to be fighting is initiated at time T if one of these persons is *active*, the other one is not *inactive* and their distance is smaller than a particular threshold. Rule (4) states that *fighting* is initiated when one of the two persons moves *abruptly*, the other is not *inactive*, and the two persons are sufficiently close. Finally, Rules (5) and (6) state that *fighting* is terminated if one of the people walks or runs away from the other.

CAVIAR contains noisy data mainly due to human errors in the annotation process (List et al, 2005; Artikis et al, 2010b). Thus, for the experiments we manually selected a noise-free subset of CAVIAR. The data consists of 1000 examples (that is, data for 1000 distinct time points) concerning the high-level events *moving together*, *meeting* and *fighting*. These data, selected from different parts of the CAVIAR dataset, were combined into a continuous annotated stream of consecutive narrative atoms, with time ranging from 0 to 1000.

In addition to the real data, we generated synthetic data on the basis of the manually-developed CAVIAR event definitions described in (Artikis et al, 2010b). In particular, streams of low-level events (narrative) concerning four different persons were created randomly and were then classified using the rules of (Artikis et al, 2010b). The final dataset was obtained by generating negative supervision via the closed world assumption and appropriately pairing the supervision with the narrative. The generated data consists of approximately 10^5 examples (that is, narrative and observations for 10^5 time points), which amounts approximately to 100 MB of data.

The synthetic data is much more complex than the real CAVIAR data. This is due to two main reasons: First, the synthetic data includes significantly more initiations and terminations of a high-level event, thus much larger learning effort is required to explain it. Second, in the synthetic dataset more than one high-level event may be initiated or terminated at the same time point. In contrast, in the real CAVIAR data only a single high-level event is initiated or terminated at some time point.

Real CAVIAR data	ILED			XHAIL
	$G = 10$	$G = 50$	$G = 100$	$G = 900$
Training Time (secs)	34.15 (± 6.87)	23.04 (± 13.50)	286.74 (± 98.87)	1560.88 (± 4.24)
Revisions	11.2 (± 3.05)	9.1 (± 0.32)	5.2 (± 2.1)	—
Hypothesis size	17.82 (± 2.18)	17.54 (± 1.5)	17.5 (± 1.43)	15 (± 0.067)
Precision	98.713 (± 0.052)	99.767 (± 0.038)	99.971 (± 0.041)	99.973 (± 0.028)
Recall	99.789 (± 0.083)	99.845 (± 0.32)	99.988 (± 0.021)	99.992 (± 0.305)
Synthetic CAVIAR data	$G = 10$	$G = 50$	$G = 100$	$G = 1000$
Training Time	38.92 (± 9.15)	33.87 (± 9.74)	468 (± 102.62)	21429 (± 342.87)
Revisions	28.7 (± 9.34)	15.4 (± 7.5)	12.2 (± 6.23)	—
Hypothesis size	143.52 (± 19.14)	138.46 (± 22.7)	126.43 (± 15.8)	118.18 (± 14.48)
Precision	55.713 (± 0.781)	57.613 (± 0.883)	63.236 (± 0.536)	63.822 (± 0.733)
Recall	68.213 (± 0.873)	71.813 (± 0.756)	71.997 (± 0.518)	71.918 (± 0.918)

Table 5: Comparative performances of ILED and XHAIL on a selected subset of the CAVIAR dataset and on synthetic data. G is the granularity (i.e number of examples) of the windows.

Part of our experimental evaluation aimed to compare ILED with XHAIL. To achieve this aim we had to implement XHAIL because the original implementation is not publicly available. All experiments were conducted on a dual core 3.2 GHz machine with 4 GB of RAM, running Ubuntu Linux 12.04. The algorithms were implemented in Java 7, using the Clingo⁴ Answer Set Solver (Gebser et al, 2012) as the main reasoning component, and a MongoDB⁵ NoSQL database for the historical memory of the examples.

7.2 Experiment 1

The purpose of this experiment was to assess whether ILED can efficiently generate hypotheses comparable in size and predictive quality to those of XHAIL. To this end, we compared both systems on real and synthetic data using 10-fold cross validation with replacement. For the real data, 90% of randomly selected examples, from the total of 1000 were used for training, while the remaining 10% was retained for testing. At each run, the training data were presented to ILED in example windows of sizes 10, 50, 100. The data were presented in one batch to XHAIL. For the synthetic data, 1000 examples were randomly sampled at each run from the dataset for training, while the remaining data were retained for testing. The data sets were presented to ILED in windows of sizes 10, 50, 100 and to XHAIL in one batch. Table 5 presents the experiment results.

Training times are significantly higher for XHAIL, as can be seen in Table 5. This is due to the increased complexity of generalizing Kernel Sets that account for the whole set of the presented examples at once. These Kernel Sets consisted, on average, of 30 to 35 16-literal clauses, in the case of the real data, and 60 to 70 16-literal clauses in the case of the synthetic data. In contrast, ILED had to deal with much smaller structures (Kernel Sets or support sets). The complexity of abductive search affects ILED as well, as the size of the input windows grows. It handles the learning task relatively easy (in approximately 30 seconds) when the examples are presented in windows of 50 examples, but the training time increases almost 15 times if the window size is doubled.

⁴ <http://potassco.sourceforge.net/>

⁵ <http://www.mongodb.org/>

Concerning the size of the outcome hypothesis, the results show that in the case of real CAVIAR data, the hypotheses constructed by ILED are comparable in size with an optimal hypothesis returned by XHAIL. In the case of synthetic data, hypotheses returned by both XHAIL and ILED were significantly more complex. Note that for ILED the hypothesis size decreases as the window size increases. This is reflected in the number of revisions that ILED performs, which is significantly smaller when the input comes in larger batches of examples. In principle, the richer the input, the better the hypothesis that is initially acquired, and consequently, the less the need for revisions in response to new training instances. There is generally a trade-off between the window size (thus the complexity of the abductive search) and the number of revisions. A small number of revisions on complex data (i.e. larger windows) may have a greater total cost in terms of training time, as compared to a greater number of revisions on simpler data (i.e. smaller windows). For example, in the case of window size 100 for the real CAVIAR data, ILED performs 5 revisions on average and requires significantly more time than in the case of a window size 50, where it performs 9 revisions on average. On the other hand, training times for windows of size 50 are slightly better than those obtained when the examples were presented in smaller windows of size 10. In this case, the “unit costs” of performing revisions w.r.t a single window are comparable between windows of size 10 and 50. Thus the overall cost in terms of training time is determined by the total number of revisions, which is greater in case of window size 10.

Concerning predictive quality, the results indicate that ILED’s precision and recall scores are comparable to those of XHAIL, with the latter being slightly better. For larger input windows precision and recall are almost the same as these of XHAIL. This is because ILED produces better hypotheses from larger input windows. Precision and recall are smaller in the case of synthetic data, because as mentioned in Section 7.1, in the case of synthetic data the testing set was much larger and complex than in the case of real data.

7.3 Experiment 2

The purpose of this experiment was to assess the scalability of the proposed method. The experimental setting was as follows: Sets of examples of varying sizes were randomly sampled from the synthetic dataset. Each such example set was used as a training set in order for ILED to acquire an initial hypothesis. Then a new window which did not satisfy the hypothesis at hand was randomly selected and presented to ILED, which subsequently revised the initial hypothesis in order to account for both the historical memory (the initial training set) and the new evidence. The aim was to assess ILED’s performance for different sizes of the historical memory and the new incoming window. For historical memories ranging from 10^3 to 10^5 examples, a new training window of size 10, 50 and 100 was selected from the whole dataset. The process was repeated ten times for each different combination of historical memory and new window sizes. Figure 2 presents the average revision times. The revision times for new window sizes of 10 and 50 examples are very close and therefore omitted from Figure 2 to avoid clutter. The results indicate that revision time grows polynomially in the size of the historical memory.

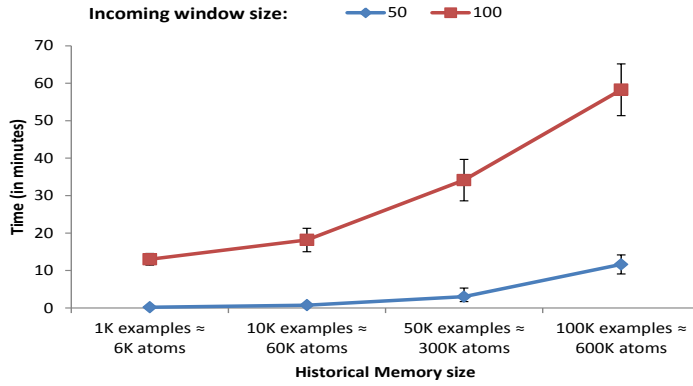


Fig. 2: Average times needed for ILED to revise an initial hypothesis in the face of new evidence presented in windows of size 10, 50 and 100 examples. The initial hypothesis was obtained from a training set of varying size (1K, 10K, 50K and 100K examples) which subsequently served as the historical memory. Average revision times for input data of size 10 and 50 are very close and presented with a single line.

8 Related work

A thorough review on the drawbacks of state-of-the-art ILP systems with respect to non-monotonic domains, as well as the deficiencies of existing approaches to learning Event Calculus programs can be found in (Ray, 2009; Sakama, 2005, 2001; Otero, 2001, 2003). The main obstacle, common to many learners which combine ILP with some form of abduction, like PROGOL5 (Muggleton and Bryant, 2000), ALECTO (Moyle, 2003), HAIL (Ray et al, 2003) and IMPARO (Kimber et al, 2009) is that they cannot perform abduction through negation and are thus essentially limited to Observational Predicate Learning.

TAL (Corapi et al, 2010) is a top-down non-monotonic learner which is able to solve the same class of problems as XHAIL. It obtains a top theory by appropriately mapping the ILP problem at hand to a corresponding ALP instance so that solutions to the latter may be translated to solutions for the initial ILP problem. Recently, the main ideas behind TAL have been employed in the ASPAL system (Corapi et al, 2011b), an inductive learner which relies on Answer Set Programming as a unifying abductive-inductive framework. ASPAL obtains a top theory of *skeleton rules* by forming all possible clause structures that may be formed from the input mode declarations. Each such structure is complemented by the addition of a set of properly formed abducible predicates. Abductive reasoning on a proper *meta-level* representation of the original ILP problem, returns a set of such abducibles, which, due to their construction, allow to hypothesize on how variables and constants in the skeleton rules are linked together. Thus, the abduced atoms are prescriptions on how variable and constant terms in the original skeleton rules should be handled in order to obtain a hypothesis. This way, ASPAL may induce all possible hypotheses w.r.t to a certain ILP problem, as well as optimal ones, by computing minimal sets of abducibles.

In the non-monotonic setting, traditional ILP approaches that cover the examples sequentially cannot ensure soundness and completeness (Sakama, 2005). To deal with this issue, non-monotonic learners like XHAIL, TAL and ASPAL gener-

alize all available examples in one go. The disadvantage of this approach, however, is poor scalability. A recent advancement which addresses the issue of scalability in non-monotonic ILP is presented in (Athakravi et al, 2013). This approach combines the top-down, meta-level learning of TAL and ASPAL, with theory revision as “non-monotonic ILP” (Corapi et al, 2008), to address the “grounding bottleneck” in ASPAL’s functionality. The top theory derived by ASPAL, as a starting point for its search, is based on combinations of the available mode declarations and it grows exponentially with the length of its clauses. Thus, obtaining a ground program from this top theory is often very expensive and can cause a learning task to become intractable (Athakravi et al, 2013). RASPAL, the system proposed in (Athakravi et al, 2013), addresses this issue by imposing bounds on the length of the top theory. Partial hypotheses of specified clause length are iteratively obtained in a refinement loop. At each iteration of this loop, the partial hypothesis obtained from the previous refinement step is further refined using theory revision as described in (Corapi et al, 2008). The process continues until a complete and consistent hypothesis is obtained. The authors show that this approach results in shorter ground programs and is ensured to derive a complete and consistent hypothesis, if one is derivable from the input data. An important difference between RASPAL and our approach is that the former addresses scalability as related to application domains, which may require a complex language bias, while our approach aims to scale to potentially simpler, but massive volumes of sequential data, typical in temporal applications.

TAL, ASPAL and RASPAL are top-down learners. In the work presented here, XHAIL, being a bottom-up non-monotonic learning system was the natural choice to use as the basis of our approach, since we intended to provide a clause refinement search bias by means of most-specific clauses, as in (Duboc et al, 2009). In that work, the Theory Revision system FORTE (Richards and Mooney., 1995) is enhanced by porting PROGOL’s bottom set construction routine to its functionality, towards a more efficient refinement operator. The resulting system, FORTE_MBC, works as follows: When a clause C must be refined, FORTE_MBC uses mode declarations and an inverse entailment search in the background knowledge to construct a bottom clause from a positive example covered by C . It then searches for antecedents within the bottom clause. As in the case of ILED, the constrained search space results in a more efficient clause refinement process. However FORTE_MBC (like FORTE itself) learns horn theories and does not support non-Observational Predicate Learning, thus it cannot be used for the revision of Event Calculus programs. In addition, it cannot operate on an empty hypothesis (i.e it cannot induce a hypothesis from scratch), but it can only revise a given hypothesis. Another important difference between FORTE_MBC and the work presented here is the way that the former handles a potential incompleteness which may result from the specialization of a clause. In particular, once a clause is specialized, FORTE_MBC checks again the whole database of examples. If some positive examples have become unprovable due to the specialization, FORTE_MBC picks a different positive example covered by the initial, inconsistent clause C , constructs a new bottom clause and searches for a new specialization of clause C . The process continues until the original coverage in the example database is restored. In contrast, by means of the support set, the specializations performed by ILED preserve prior coverage in the historical memory, thus saving a lot of inference effort.

As mentioned in (Duboc et al, 2009), there is a renewed interest in scaling Theory Revision systems and applications in the last few years, due to the availability of large-scale domain knowledge in various scientific disciplines (Dietterich et al, 2008; Muggleton et al, 2012a). Temporal and stream data are no exception and there is a need for scalable Theory Revision techniques in event-based domains. However, most Theory Revision systems, such as the systems described in (Richards and Mooney, 1991; Quinlan, 1990; Wogulis and Pazzani, 1993) limit their applicability to Horn theories.

A well-known theory revision system is INTHELEX (Esposito et al, 2000). It is a fully incremental system that learns/revises Datalog theories and has been used in the study of several aspects of incremental learning. In particular, order effects in some simple learning tasks with ILP are discussed in (Mauro et al, 2004, 2005) and concept drift in (Esposito et al, 2004). In (Biba et al, 2008) the authors present an approach towards scaling INTHELEX, which is related to the one presented here. In contrast to most ILP systems that keep all examples in the main memory, (Biba et al, 2008) follows an external memory implementation, which is the approach adopted also in this paper. Additionally, in that work the authors associate clauses in the theory at hand with examples they cover, via a relational schema. Thus, when a clause is refined, only the examples that were previously covered by this clause are checked. Similarly, when a clause is generalized, only the negative examples are checked again. Finally, the scalable version of INTHELEX presented in (Biba et al, 2008) maintains alternative versions of the hypothesis at each step, allowing to backtrack to previous states. In addition, it keeps in memory several statistics related to the examples that the system has already seen, such as the number of refinements that each example has caused, a “refinement history” of each clause etc.

On the other hand, INTHELEX has some limitations that make it inappropriate for inducing/revising Event Calculus programs for event recognition applications. First, the restriction of its input language to Datalog limits its applicability to richer, relational event domains. For instance, fluents and events that express complex relations between entities are hard to be expressed with INTHELEX. Second, use of background knowledge is limited, excluding for instance auxiliary clauses that may be used for spatio-temporal reasoning during learning time. Third, although INTHELEX uses abduction for the completion of imperfect input data, it relies on Observational Predicate Learning, meaning that it is not able to reason with predicates which are not directly observable in the examples. Therefore it cannot be used for learning event definitions.

9 Conclusions

We presented an incremental version of the ILP system XHAIL, applicable to the acquisition of event-based knowledge in the form of Event Calculus theories. Although the proposed approach is oriented towards temporal data, however it may also facilitate learning in many applications that involve sequential data with a time-like structure. The presented methodology is implemented as a full memory system, allowing revisions which are correct with respect to all training instances seen so far. The main contribution is efficient clause refinement, based on abductive reasoning on a properly constrained antecedent search space. This search

results from a bottom clause-based compressive summarization of the parts of the historical memory which are relevant to the example coverage of the hypothesis at hand. The overall approach leads to an efficient and scalable refinement operator, under certain locality assumptions. We evaluated our system on data from a real benchmark dataset, as well as large-scale synthetic data. The results indicate that our approach is significantly more efficient than XHAIL, without compromising predictive accuracy, and scales adequately to large data volumes. In addition, we identified various directions for future work, including mechanisms for handling noise and concept drift and introduction of non-monotonic refinement operators towards the induction of more compressive hypotheses.

References

- Ade H, Denecker M (1995) AILP: Abductive inductive logic programming. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence
- Alrajeh D, Kramer J, Russo A, Uchitel (2009) Learning perational requirements from goal models. In: 31st International Conference on Software Engineering
- Alrajeh D, Kramer J, Russo A, Uchitel S (2010) Deriving non-zeno behaviour models from goal models using ilp. *Formal Aspects of Computing* 22(3-4):217–241
- Alrajeh D, Kramer J, Russo A, Uchitel S (2011) An inductive approach for modal transition system refinement. In: International Conference on Logic Programming
- Alrajeh D, Kramer J, Russo A, Uchitel (2012) Learning from vacuously satisfiable scenario-based specifications. In: 15th International Conference on Fundamental Approaches to Software Engineering (FASE)
- Artikis A, Paliouras G, Portet F, Skarlatidis A (2010a) Logic-based representation, reasoning and machine learning for event recognition. In: Proceeding of Destrubuted Event Based Systems (DEBS), pp 282–293
- Artikis A, Skarlatidis A, Paliouras G (2010b) Behaviour recognition from video content: A logic programming approach. *International Journal on Artificial Intelligence Tools* 19(2):193–209
- Athakravi D, Corapi D, Broda K, Russo A (2013) Learning through hypothesis refinement using answer set programming. In: Proceedings of the 23rd International Conference of Inductive Logic Programming (ILP 2013)
- Badea L (2001) A refinement operator for theories. In: Inductive Logic Programming, Springer, pp 1–14
- Biba M, Basile TMA, Ferilli S, Esposito F (2008) Improving scalability in ilp incremental systems
- Blockeel H, Raedt LD, Jacobs N, Demoen B (1999) Scaling up inductive logic programming by learning from interpretations. *Data Mining and Knowledge Discovery* 3:59–93
- Cattafi M, Lamma E, Riguzzi F, Storari S (2010) Incremental declarative process mining. *Smart Information and Knowledge Management* pp 103–127
- Clark KL (1977) Negation as failure. *Logic and Data Bases* pp 293–322
- Corapi D, Ray O, Russo A, Bandara A, Lupu E (2008) Learning rules from user behaviour. In: Second International Workshop on the Induction of Process Models

- Corapi D, Russo A, Lupu EC (2010) Inductive logic programming as abductive search. In: Technical Communications of the 26th International Conference on Logic Programming (ICLP)
- Corapi D, De Vos M, Padget J, Russo A, Satoh K (2011a) Norm refinement and design through inductive learning. In: Coordination, Organizations, Institutions, and Norms in Agent Systems VI, Springer, pp 77–94
- Corapi D, Russo A, Lupu E (2011b) Inductive logic programming in answer set programming. In: ILP
- Denecker M, Kakas A (2002) Abduction in logic programming. *Computational Logic: Logic Programming and Beyond* 2407:402–437
- Dietterich TG, Domingos P, Getoor L, Muggleton S, Tadepalli P (2008) Structured machine learning: the next ten years. *Machine Learning* 73:3–23
- Dubba K, Bhatt M, Dylla F, Cohn A, Hogg D (2011) Interleaved inductive-abductive reasoning for learning event-based activity models. In: Inductive Logic Programming - 21st International Conference, ILP 2011
- Duboc AL, Paes A, Zaverucha G (2009) Using the bottom clause and mode declarations in FOL theory revision from examples. *Machine Learning* 76(1):73–107
- Eshghi K, Kowalski R (1989) Abduction compared with negation by failure. In: 6th International Conference on Logic Programming
- Esposito F, Semeraro G, Fanizzi N, Ferilli S (2000) Multistrategy theory revision: Induction and abduction in inthelex. *Machine Learning* 28(1-2):133–156
- Esposito F, Ferilli S, Fanizzi N, Basile TMA, Mauro ND (2004) Incremental learning and concept drift in inthelex. *Intelligent Data Analysis* 8(3):213–237
- Etzion O, Niblett P (2010) Event processing in action. Manning Publications Co.
- Gebser M, Kaminski R, Kaufmann B, Schaub T (2012) Answer set solving in practice. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 6(3):1–238
- Kakas A, Mancarella P (1990) Generalised stable models: A semantics for abduction. In: ninth European Conference on Artificial Intelligence (ECAI-90), pp 385–391
- Kakas A, Kowalski R, Toni F (1993) Abductive logic programming. *Journal of Logic and Computation* 2:719–770
- Kimber T, Broda K, Russo A (2009) Induction on failure: Learning connected horn theories. *Logic Programming and Nonmonotonic Reasoning, Lecture Notes in Computer Science* 5753:169–181
- Kowalski R, Sergot M (1986) A logic-based calculus of events. *New Generation Computing* 4(1):6796
- Kuzelka O, Zelezny F (2008) A restarted strategy for ecient subsumption testing. *Fundamenta Informaticae* 89
- Lavrac N, Dzeroski S (1993) Inductive Logic Programming: Techniques and Applications. Routledge
- List T, Bins J, Vazquez J, Fisher RB (2005) Performance evaluating the evaluator. In: Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on, IEEE, pp 129–136
- Lloyd J (1987) Foundations of Logic Programming. Springer
- Luckham D (2001) The Power of Events: An Introduction to Complex Event Processing in Distributed Enterprise Systems. Addison-Wesley Longman Publishing Co., Inc

- Luckham D, Schulte R (2008) Event processing glossary version 1.1. Event Processing Technical Society
- Maggi FM, Corapi D, Russo A, Lupu E, Visaggio G (2011) Revising process models through inductive learning. In: Business Process Management Workshops, Springer, pp 182–193
- Maloberti J, Sebag M (2004) Fast theta-subsumption with constraint satisfaction algorithms. *Machine Learning* 55
- Mancarella P, Terreni G, Sadri F, Toni F, Endriss U (2009) The ciff proof procedure for abductive logic programming with constraints: Theory, implementation and experiments. *Theory and Practice of Logic Programming (TPLP)* 9(6):691–750
- Marek V, Truszczyński M (1999) *The Logic Programming Paradigm*, Springer Verlag, chap Stable Models and an Alternative Logic Programming Paradigm, pp 375–398
- Mauro ND, Esposito F, Ferilli S, Basile TM (2004) A backtracking strategy for order-independent incremental learning. In: *Proceedings of ECAI04*
- Mauro ND, Esposito F, Ferilli S, 110-121 TB (2005) Avoiding order effects in incremental learning. In: *AIIA 2005: Advances in Artificial Intelligence*,
- Moyle S (2003) An investigation into theory completion techniques in inductive logic. PhD thesis, University of Oxford
- Mueller E (2006) *Commonsense Reasoning*. Morgan Kaufmann
- Mueller E (2008) Event calculus. *Handbook of Knowledge Representation 3 of FAI*:671–708
- Muggleton S (1995) Inverse entailment and prolog. *New Generation Computing* 13(3&4):245–286
- Muggleton S, Bryant C (2000) Theory completion using inverse entailment. In: *International Conference on Inductive Logic Programming*, pp 130–146
- Muggleton S, Raedt LD (1994) Inductive logic programming: Theory and methods. *Journal of Logic Programming* 19/20:629679
- Muggleton S, Paes A, Costa VS, Zaverucha G (2012a) Chess revision: acquiring the rules of chess variants through fol theory revision from examples. In: *Inductive Logic Programming*
- Muggleton S, Raedt LD, Poole D, Bratko I, Flach P, Inoue K, Srinivasan A (2012b) ILP turns 20 - biography and future challenges. *Machine Learning* 86(1):3–23
- Niemela I (1999) Logic programs with stable model semantics as a constraint programming paradigm. *Annals of Mathematics and AI* 25(3-4)
- Otero RP (2001) Induction of stable models. In: *Inductive Logic Programming*, Springer, pp 193–205
- Otero RP (2003) Induction of the effects of actions by monotonic methods. In: *Inductive Logic Programming*, Springer, pp 299–310
- Paschke A (2005) Eca-ruleml: An approach combining eca rules with temporal interval-based kr event logics and transactional update logics. Tech. rep., Technische Universität München
- Quinlan JR (1990) Learning logical definitions from relations. *Machine Learning* 5:239266
- Ray O (2006) Using abduction for induction of normal logic programs. In: *ECAI06 Workshop on Abduction and Induction in Artificial Intelligence and Scientific Modelling*
- Ray O (2009) Nonmonotonic abductive inductive learning. *Journal of Applied Logic* 7(3):329–340

- Ray O, Kakas A (2006) ProLogICA: a practical system for abductive logic programming. In: 11th International Workshop on Non-monotonic Reasoning, p 304312
- Ray O, Broda K, Russo A (2003) Hybrid abductive inductive learning: A generalisation of prolog. In: International Conference in Inductive Logic Programming (ILP), pp 311–328
- Richards B, Mooney R (1995) Automated refinement of first-order horn clause domain theories. *Machine Learning* 19(2):95–131
- Richards BL, Mooney RJ (1991) First order theory revision. In: 8th International Workshop on Machine Learning, p 447451
- Sakama C (2000) Inverse entailment in nonmonotonic logic programs. In: n Proceedings of the 10th International Conference on Inductive Logic Programming
- Sakama C (2001) Non-monotonic inductive logic programming. In: Logic Programming and Non-Monotonic Reasoning
- Sakama C (2005) Induction from answer sets in nonmonotonic logic programs. *ACM Transactions on Computational Logic* 6 (2):203231
- Santos J, Muggleton S (2010) Subsumer: A prolog theta-subsumption engine. In: Technical Communications of the 26th International Conference on Logic Programming, Leibniz International Proceedings in Informatics
- Sloman M, Lupu E (2010) Engineering policy-based ubiquitous systems. *The Computer Journal* 53(5):1113–1127
- Wogulis J, Pazzani M (1993) A methodology for evaluating theory revision systems: Results with audrey ii. In: 13th International Joint Conference in Artificial Intelligence IJCAI, pp 1128–1134
- Wrobel S (1996) First order theory refinement. In: Raedt LD (ed) *Advances in Inductive Logic Programming*, pp 14 – 33